Linguistics with CLARIN TTNWW & PaQu

Jan Odijk

LOT Winterschool

Amsterdam, 2015-01-15

Overview

- Enrich your own resources (TTNWW)
- Enrich your own resources (other applications)
- Search in your own enriched resources (AutoSearch)
- Search in your own enriched resources (PaQu)
- Conclusions

Overview

- **Enrich your own resources (TTNWW)**
- Enrich your own resources (other applications)
- Search in your own enriched resources (AutoSearch)
- Search in your own enriched resources (PaQu)
- Conclusions

TTNWW

- TTNWW
  - *TST tools voor het Nederlands als Webservices in een Workflow* (LST Tools for Dutch as web services in a work flow)
  - Goal: to make existing components for Dutch from previous projects (*inter alia* CGN and STEVIN) available as web services and combine them into work flows for use by HSS researchers with little or no technical background.

TTNWW

- Web service: software that can exchange information with other software over the internet
- (Web) application: software targeted at a specific user group with a dedicated (web) interface
- Work Flow: combinations of web services

TTNWW

- **Work Flow** describes the tasks, procedural steps, organizations or people involved, required input and output information, and tools needed for each step in a business process

TTNWW

- **Work Flow** describes the tasks, procedural steps, ~~organizations or people involved~~, required input and output information, and tools needed for each step in a business process
- (sometimes called *processing chain* in the context of software)

Existing Components

- **Text**
  - Orthographic Normalisation (TICCLops, UvT)
  - Part of Speech tagging (Frog, UvT/RUN)
  - Parsing (Alpino, RUG)
  - Named Entity Recognition (Frog, Ghent)
  - Semantic role labelling (MBSRL, UU)
  - Coreference Assignment (COREA, Antwerpen/Ghent)
- **Speech**
  - (Partial) transcription of speech (SPRAAK, KUL)

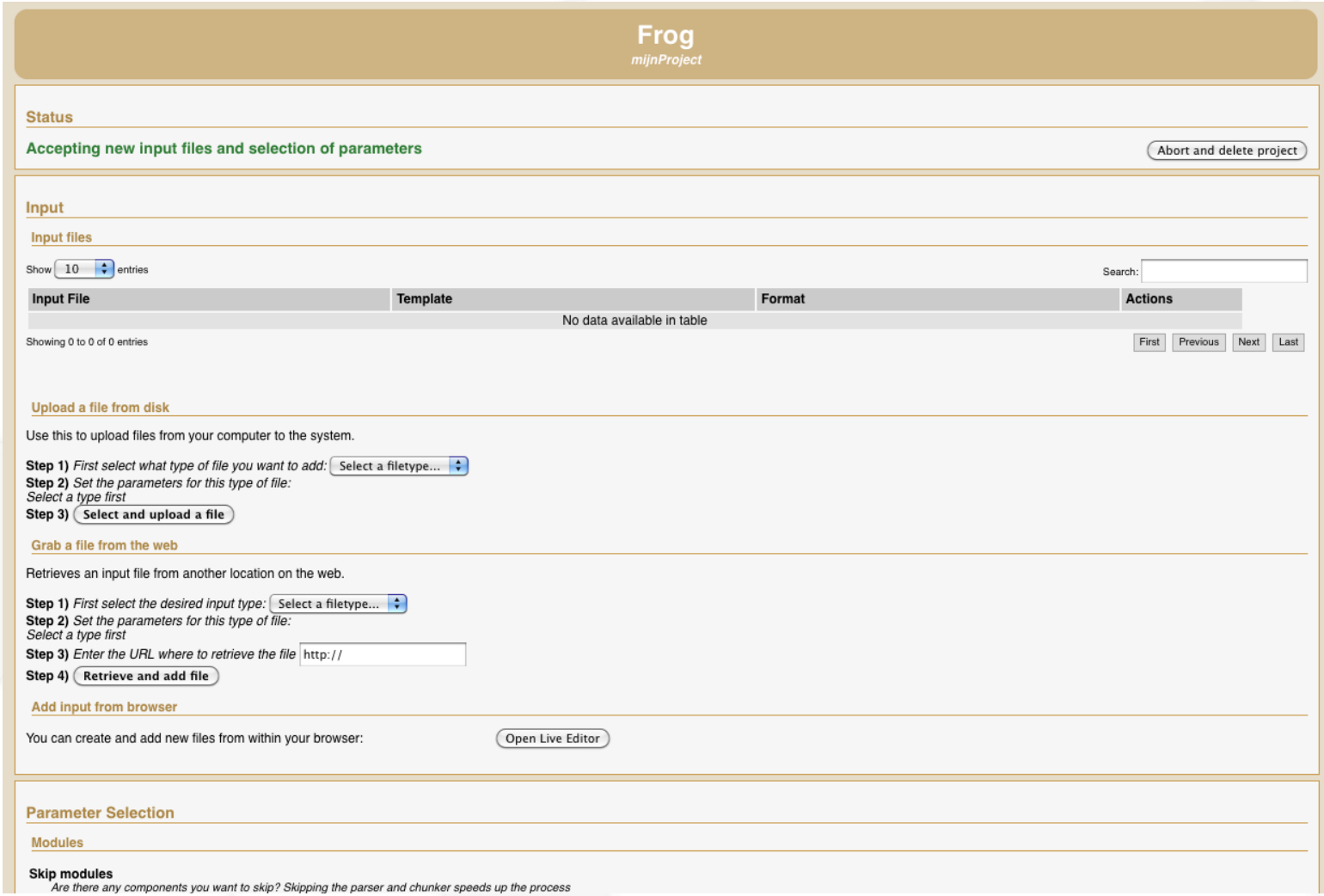
Existing Components

- For
  - Enrichment of text with corrections of OCR-input
  - Enrichment of text with linguistic annotations
  - Indexation of audio-visual files to make them searchable

Existing Components

- Before TTNWW
  - These required local installation (most under Linux only)
  - Required running by the developers
- After TTNWW
  - All available via the TTNWW web application
  - Most available individually as a web service / application

Frog Screenshot



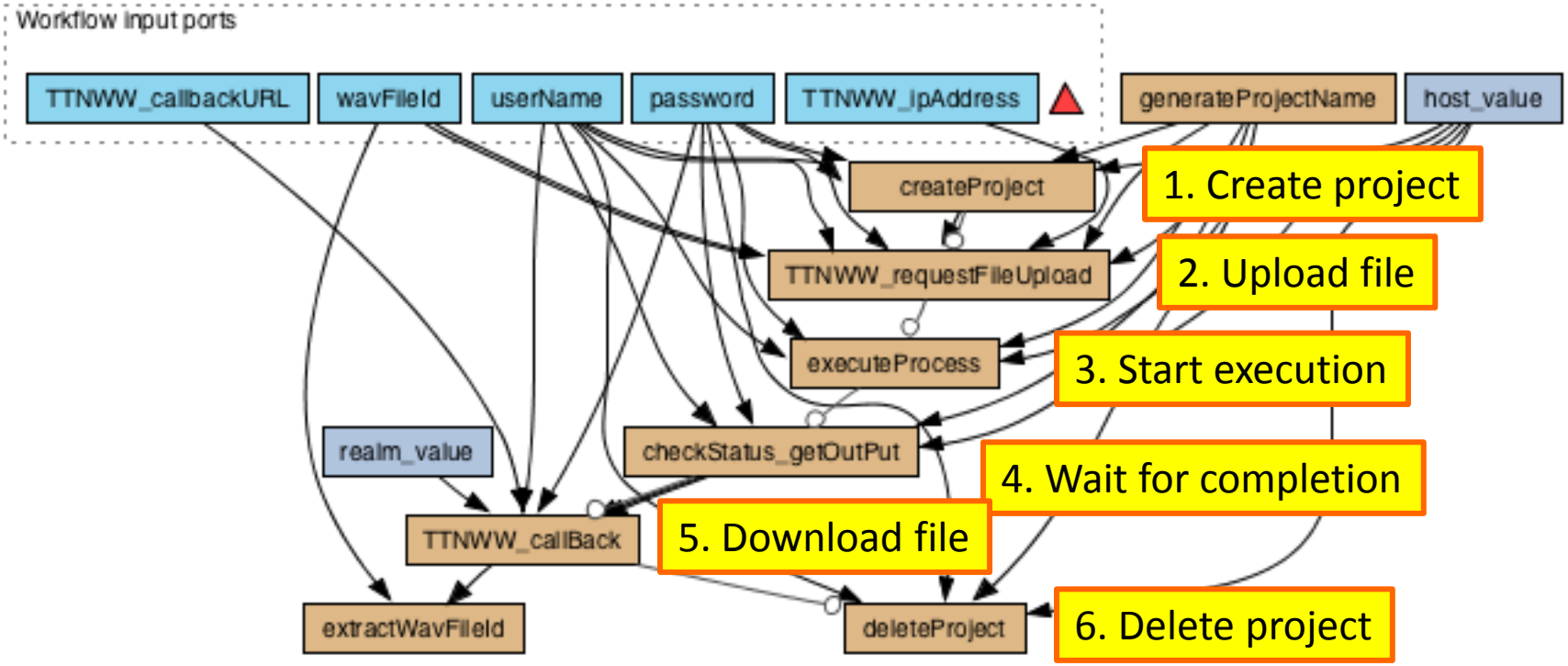
Existing Components

- Before TTNWW
  - Were command line tools
  - Working on a single system
  - No web services
- CLAM
  - Computational Linguistics Application Mediator
  - Generic wrapper to turn existing software easily into a web service
  - Most TTNWW components use CLAM

Existing Components

- **Work flows (simplified examples)**
  - Part of Speech tagging:
    - Tokenization -> sentence splitting -> Pos tagging
  - Parsing
    - Tokenization -> sentence splitting -> pos-tagging -> parsing
  - Named Entity Recognition
    - Pos-tagging -> NER
  - Semantic role labelling (MBSRL, UU)
    - Parsing -> semantic role labeling

Existing Components



Existing Components

- For TTNWW, the Taverna workbench is used
- Taverna enables creation of one's own work flows, but
- In TTNWW only fixed, predefined work flows are available
  - Making your own work flow is pretty complicated
  - The targeted audience – humanities researchers – seldom want to do this

TTNWW APPLICATION

TTNWW | TST Tools voor het Nederlands als Webservices in een Workflow

File Tools Help

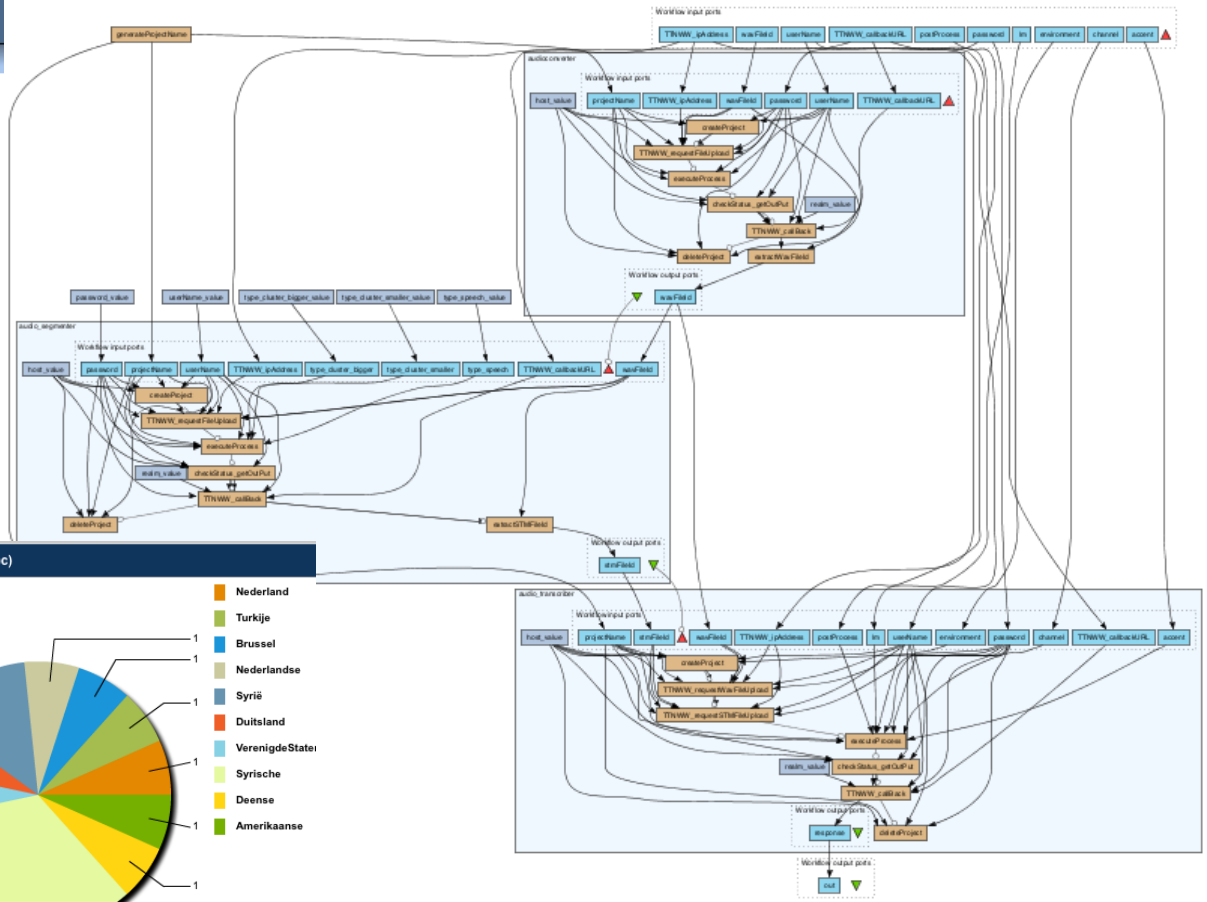
Text
Audio

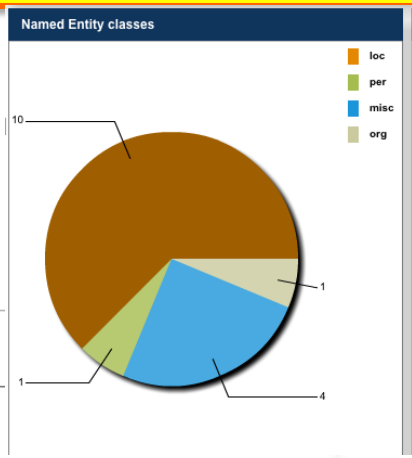
Spelling variation
Tagging
Parsing
Named Entity Recognition

Workspace Processes

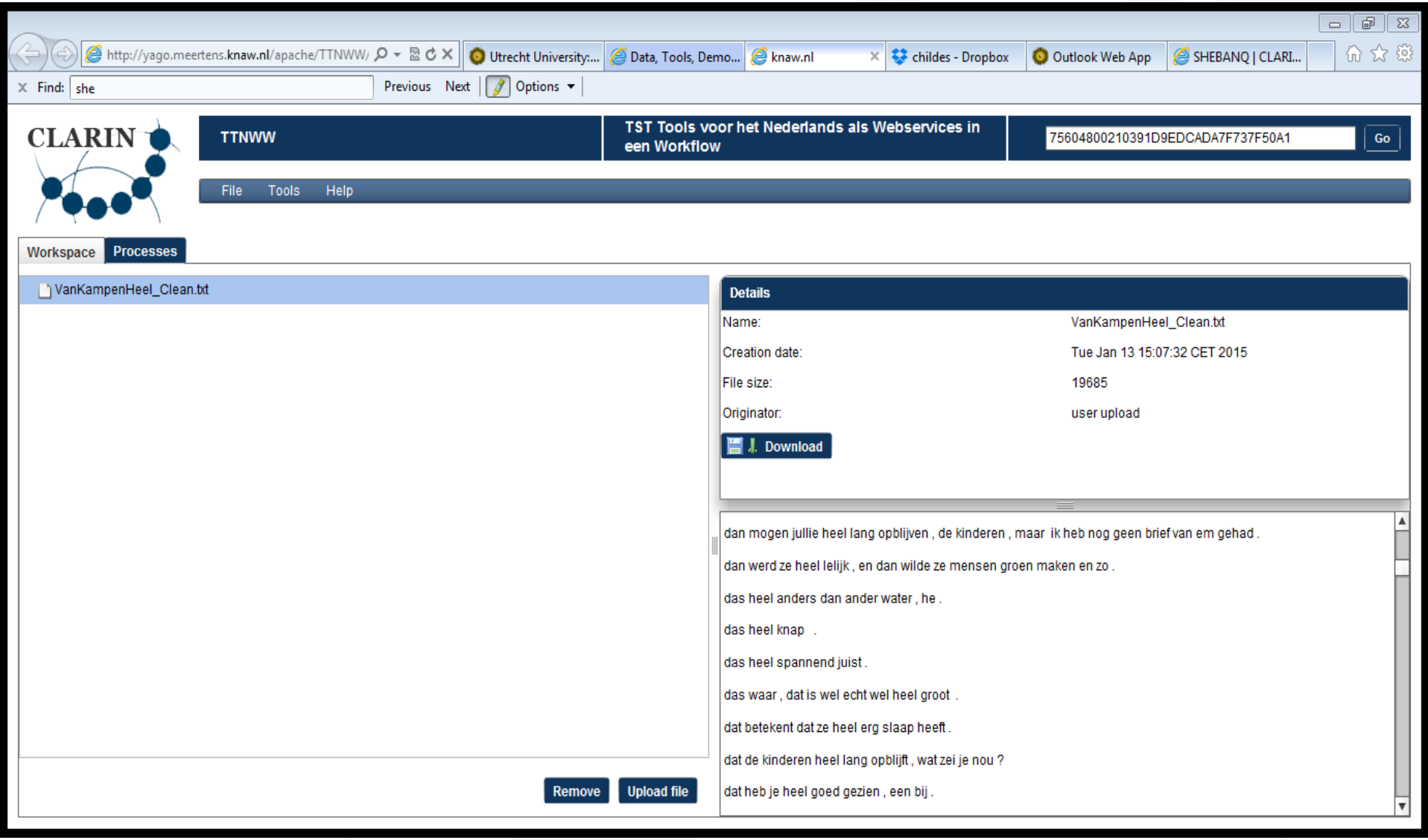
frogtest.txt
frogtest2.txt

Functionality
- *Data selection (upload/download)*
- *Selection of processes*
- *Calling processes*
- *Puts results in a **temporary** work space*
- *Only generic viewers*

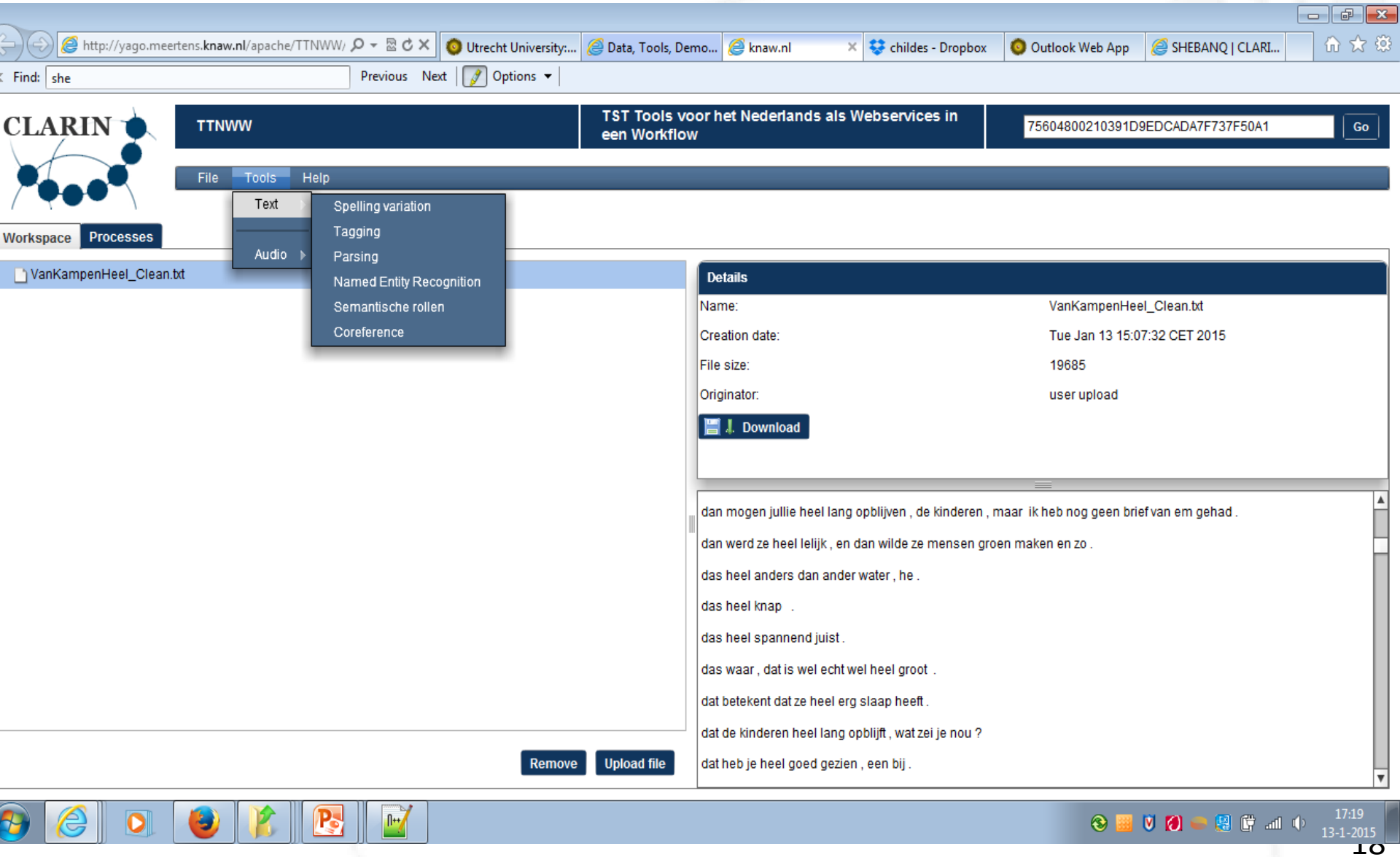




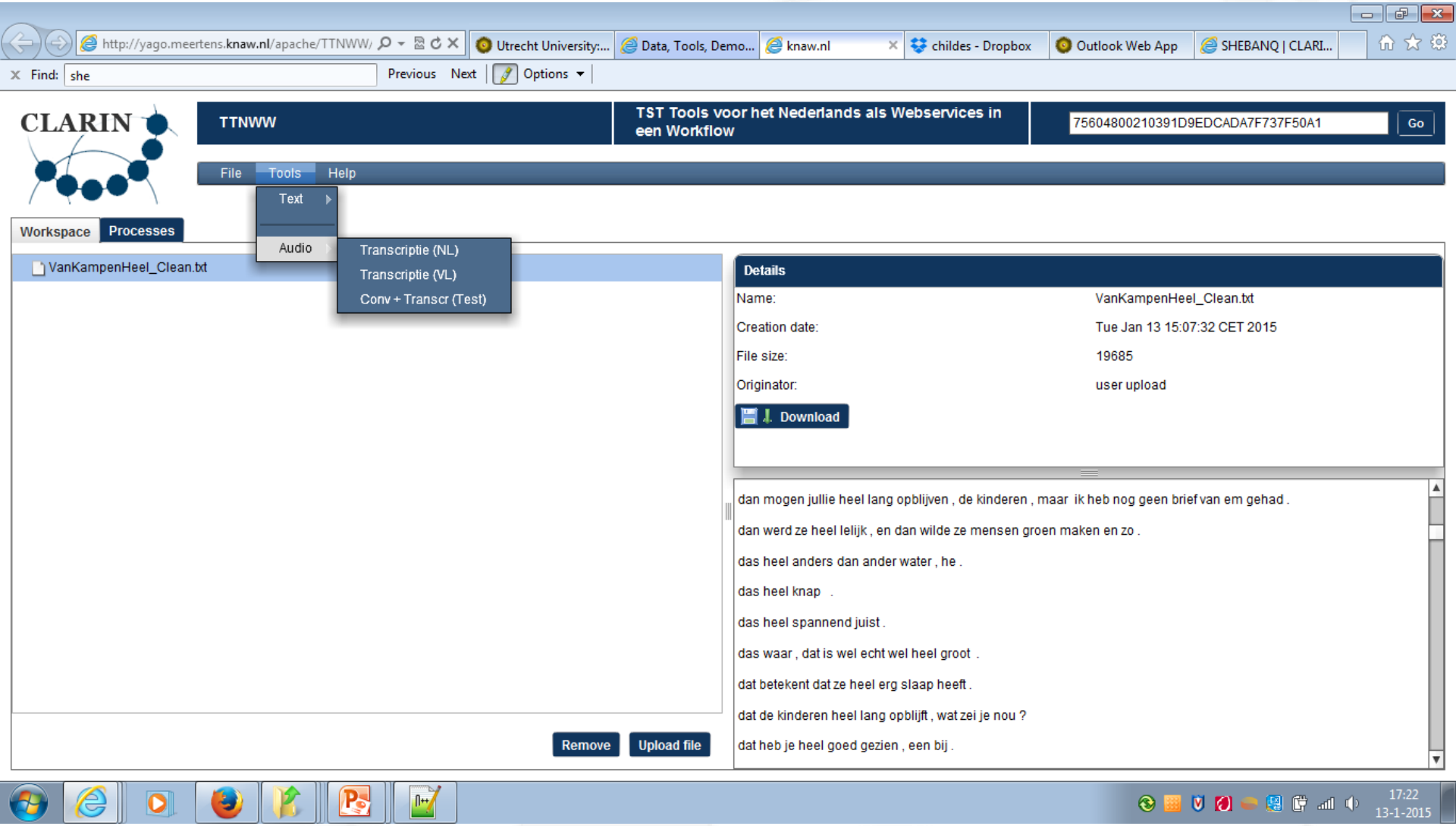
TTNWW Screen Shot



TTNWW Text


http://yago.meertens.knaw.nl/apache/TTNWW/
Utrecht University:...
Data, Tools, Demo...
knaw.nl
childes - Dropbox
Outlook Web App
SHEBANQ | CLARI...
Find: she
Previous
Next
Options
CLARIN
TTNWW
TST Tools voor het Nederlands als Webservices in een Workflow
75604800210391D9EDCADA7F737F50A1
Go
File
Tools
Help
Text
Audio
Spelling variation
Tagging
Parsing
Named Entity Recognition
Semantische rollen
Coreference
Workspace
Processes
VanKampenHeel_Clean.txt
Details
Name:
VanKampenHeel_Clean.txt
Creation date:
Tue Jan 13 15:07:32 CET 2015
File size:
19685
Originator:
user upload
Download
dan mogen jullie heel lang opblijven , de kinderen , maar ik heb nog geen brief van em gehad .
dan werd ze heel lelijk , en dan wilde ze mensen groen maken en zo .
das heel anders dan ander water , he .
das heel knap .
das heel spannend juist .
das waar , dat is wel echt wel heel groot .
dat betekent dat ze heel erg slaap heeft .
dat de kinderen heel lang opblijft , wat zei je nou ?
dat heb je heel goed gezien , een bij .
Remove
Upload file
17:19
13-1-2015

TTNWW Audio


http://yago.meertens.knaw.nl/apache/TTNWW/
Utrecht University:...
Data, Tools, Demo...
knaw.nl
childes - Dropbox
Outlook Web App
SHEBANQ | CLARI...
Find: she
Previous
Next
Options
CLARIN
TTNWW
TST Tools voor het Nederlands als Webservices in een Workflow
75604800210391D9EDCADA7F737F50A1
Go
File
Tools
Help
Text
Audio
Transcriptie (NL)
Transcriptie (VL)
Conv + Transcr (Test)
Workspace
Processes
VanKampenHeel_Clean.txt
Details
Name: VanKampenHeel_Clean.txt
Creation date: Tue Jan 13 15:07:32 CET 2015
File size: 19685
Originator: user upload
Download
dan mogen jullie heel lang opblijven , de kinderen , maar ik heb nog geen brief van em gehad .
dan werd ze heel lelijk , en dan wilde ze mensen groen maken en zo .
das heel anders dan ander water , he .
das heel knap .
das heel spannend juist .
das waar , dat is wel echt wel heel groot .
dat betekent dat ze heel erg slaap heeft .
dat de kinderen heel lang opblijft , wat zei je nou ?
dat heb je heel goed gezien , een bij .
Remove
Upload file
17:22
13-1-2015

Overview

- Enrich your own resources (TTNWW)
- ➢ **Enrich your own resources (other applications)**
- Search in your own enriched resources (AutoSearch)
- Search in your own enriched resources (PaQu)
- Conclusions

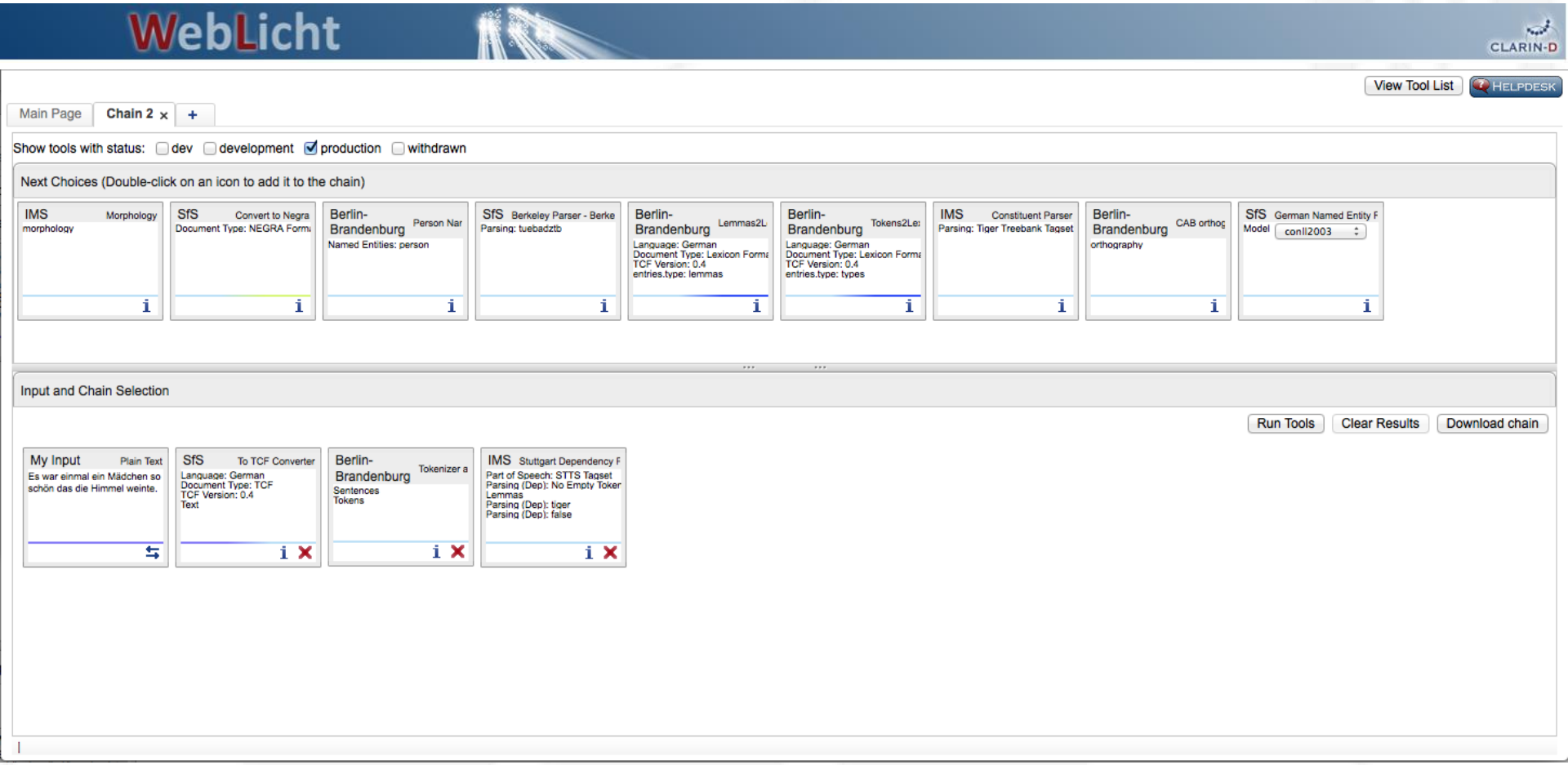
Other approaches in the '*CLARIN*' domain

- Weblicht (login with your university account)
  - *Easy Mode* lets you choose pre-defined processing chains
  - *Features Mode* allows you to select individual features, and then choose from a selection of processing chains.
  - *Advanced Mode* allows you to build customized tool chains.
- Multiple languages, mainly German

WEBLICHT D-SPIN PROJECT

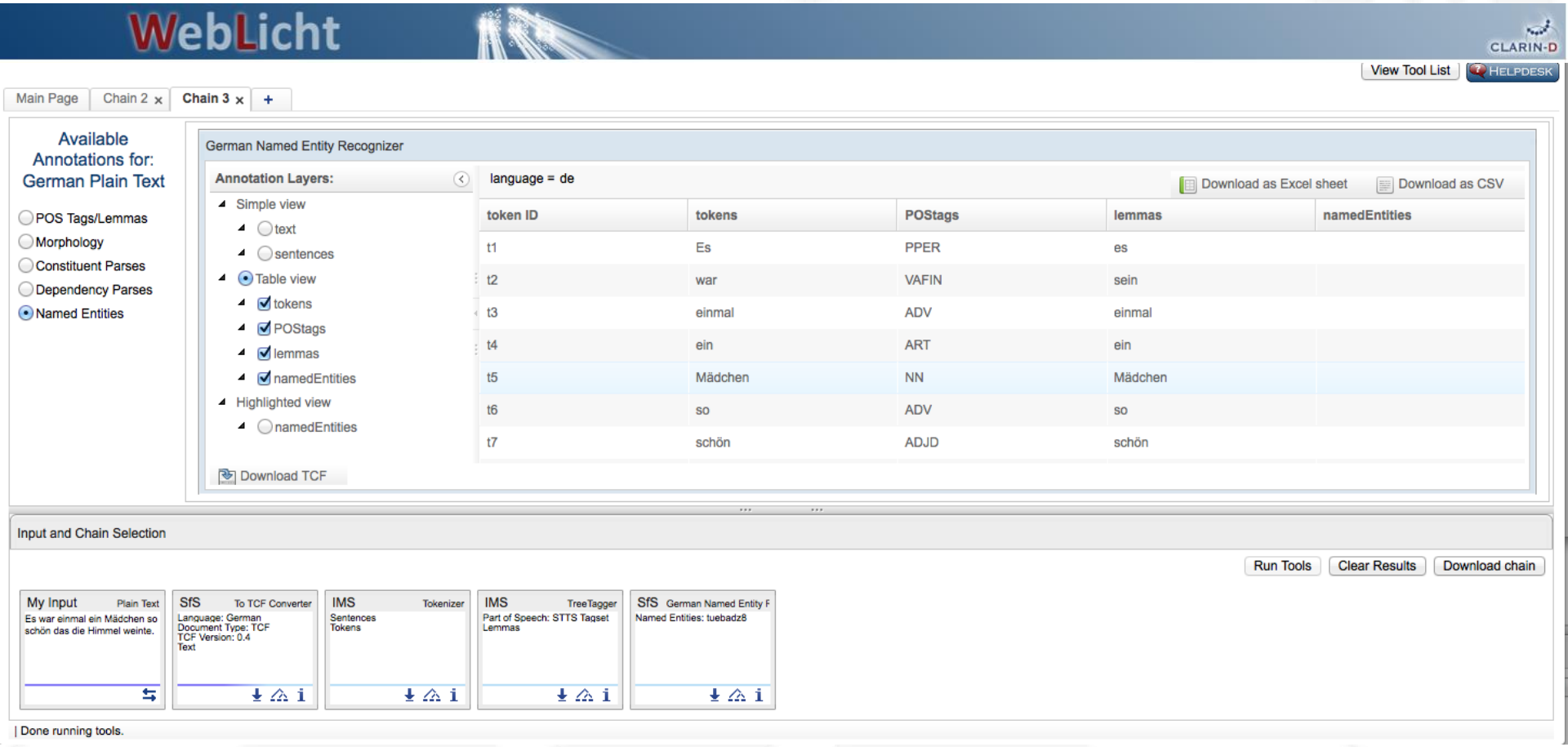
https://weblicht.sfs.uni-tuebingen.de/WebLicht-4/

WebLicht D-Spin project



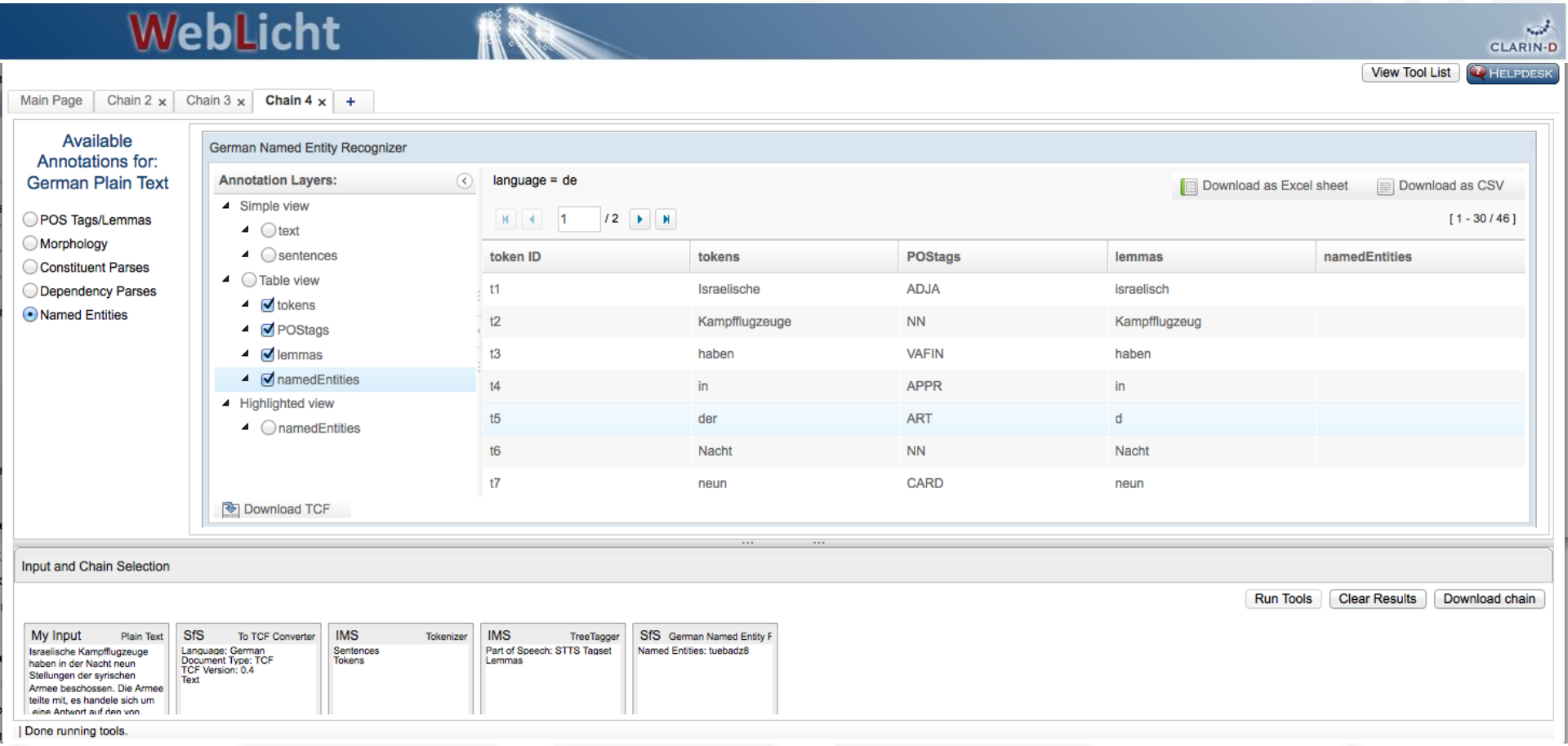
https://weblicht.sfs.uni-tuebingen.de/WebLicht-4/

WebLicht D-Spin project



https://weblicht.sfs.uni-tuebingen.de/WebLicht-4/

WebLicht D-Spin project



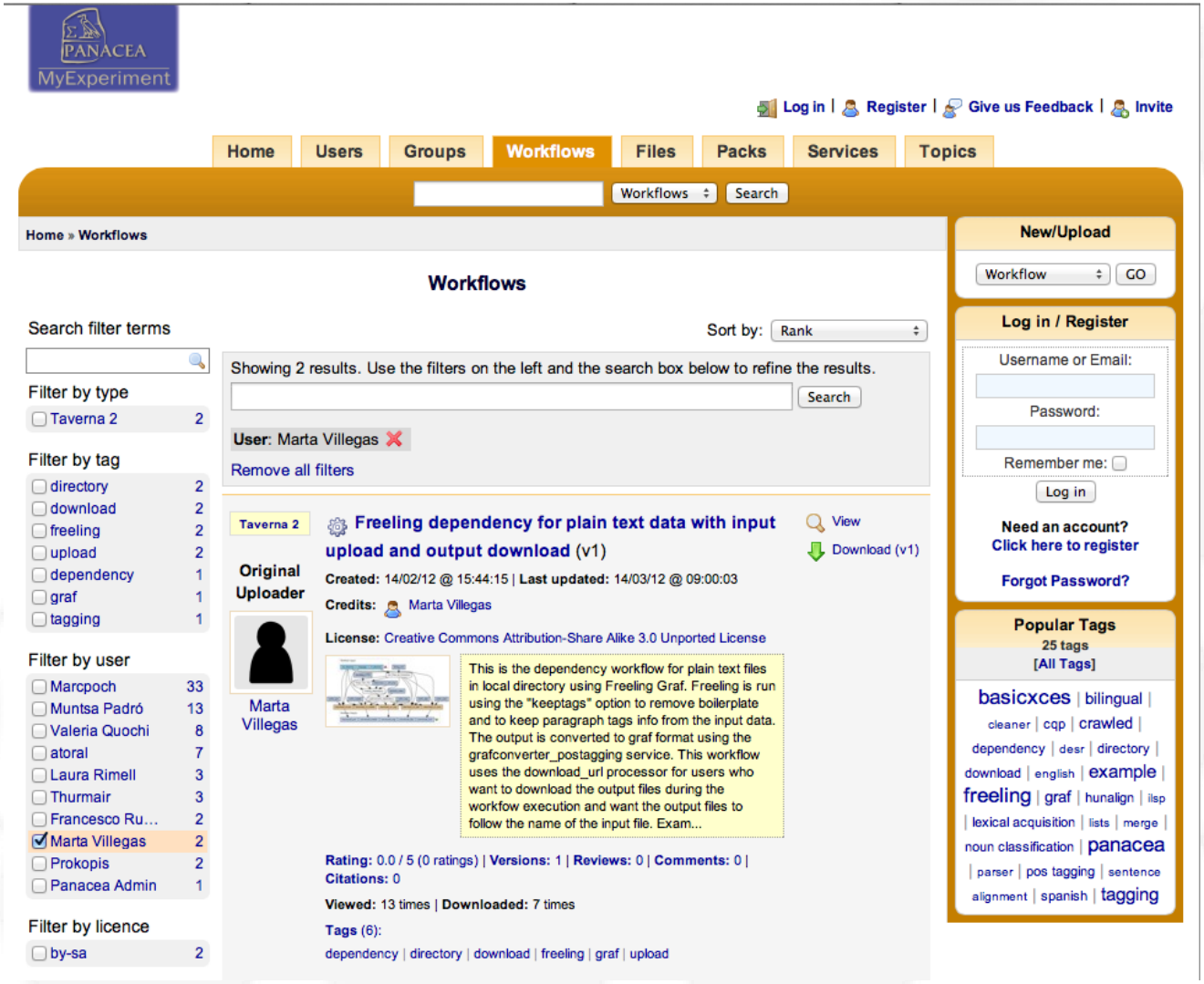
https://weblicht.sfs.uni-tuebingen.de/WebLicht-4/

Other approaches in the '*CLARIN*' domain

- Panacea
  - Taverna based workflow system
  - You can use find workflows
  - You can add your own work flows
  - You can use workflows
    - In the Taverna workbench (that you must download and install)
  - Multiple languages, incl Italian and Spanish
  - Uses Freeling (see its on-line demo)

PANACEA PROJECT

http://www.panacea-lr.eu



http://www.panacea-lr.eu

Other NLP services in the '*CLARIN*' domain

- Several services at LINDAT/CLARIN (Czech Republic), including Treex::Web (Czech, English)
- Several on-line services at CLARIN-DK, multiple languages
- Some services at CLARIN-PL (mainly for Polish)
- Language Analysis Portal by CLARINO (under development, mainly Norwegian)
- …

Overview

- Enrich your own resources (TTNWW)
- Enrich your own resources (other applications)
- ➢ **Search in your own enriched resources (AutoSearch)**
- Search in your own enriched resources (PaQu)
- Conclusions

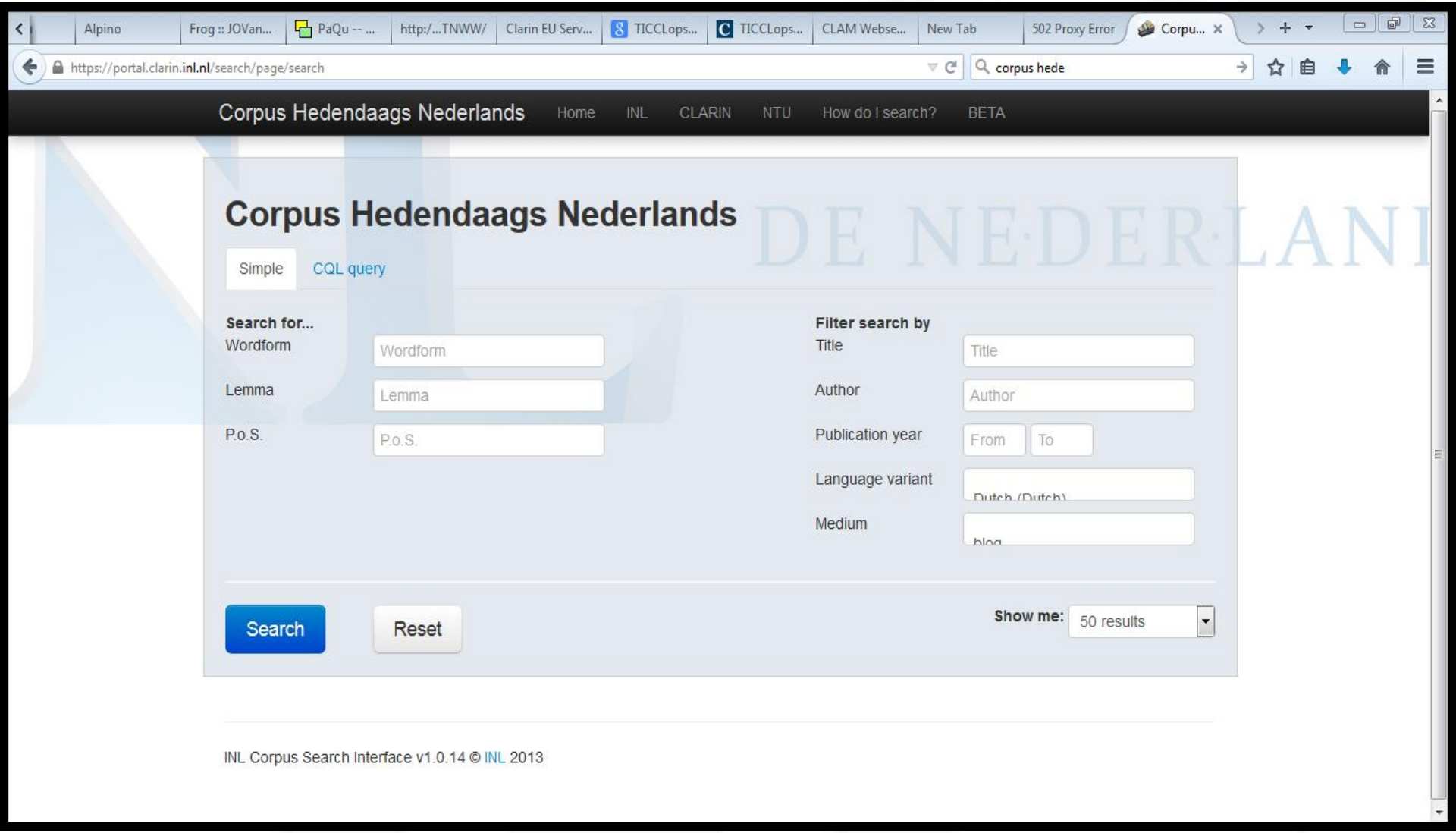
Search & Analysis

- Enriching your data is fine but
- What will you do with a (large and complex) set of XML files?
- You need options to search in the data and to analyze the data
- Two projects running in NL
  - AutoSearch (INL)
  - PaQu (Groningen University)

AutoSearch

- (under development, available March 2015)
- For pos-tagged data
- Supported formats:
  - FoLia (Frog yields Folia format)
  - TEI
- Search Interface: Corpus Hedendaags Nederlands (Corpus of Modern Dutch) interface

Corpus Hedendaags Nederlands



Overview

- Enrich your own resources (TTNWW)
- Enrich your own resources (other applications)
- Search in your own enriched resources (AutoSearch)
- ➢ **Search in your own enriched resources (PaQu)**
- Conclusions

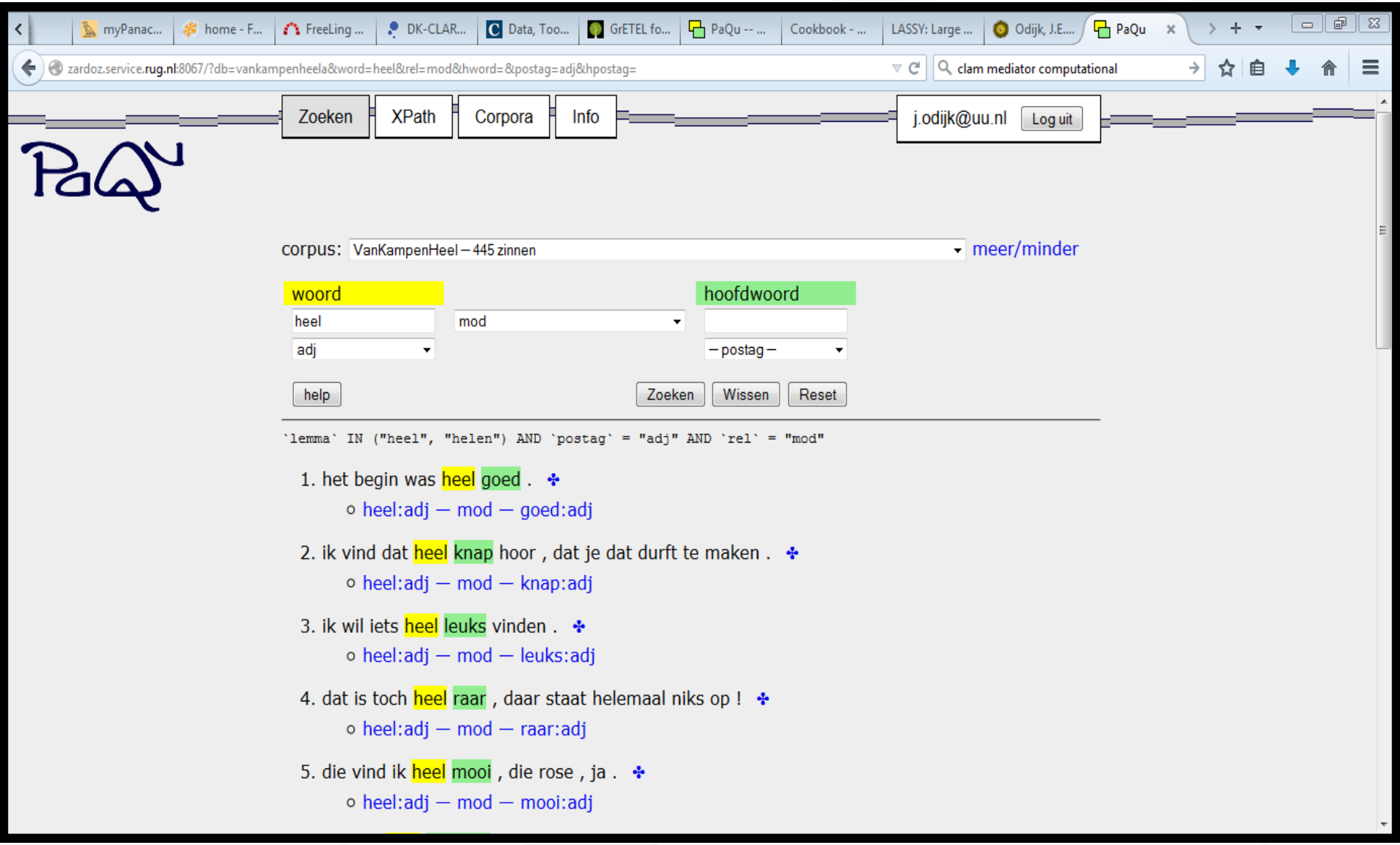
PaQu

- under development
  - available July 2015
  - Preliminary version available now!
- For fully parsed data
- Supported formats:
  - LASSY XML (yielded by TTNWW)
  - Plain text (and PaQu does the parsing)
- Search Interface: Groningen Word Relations Interface

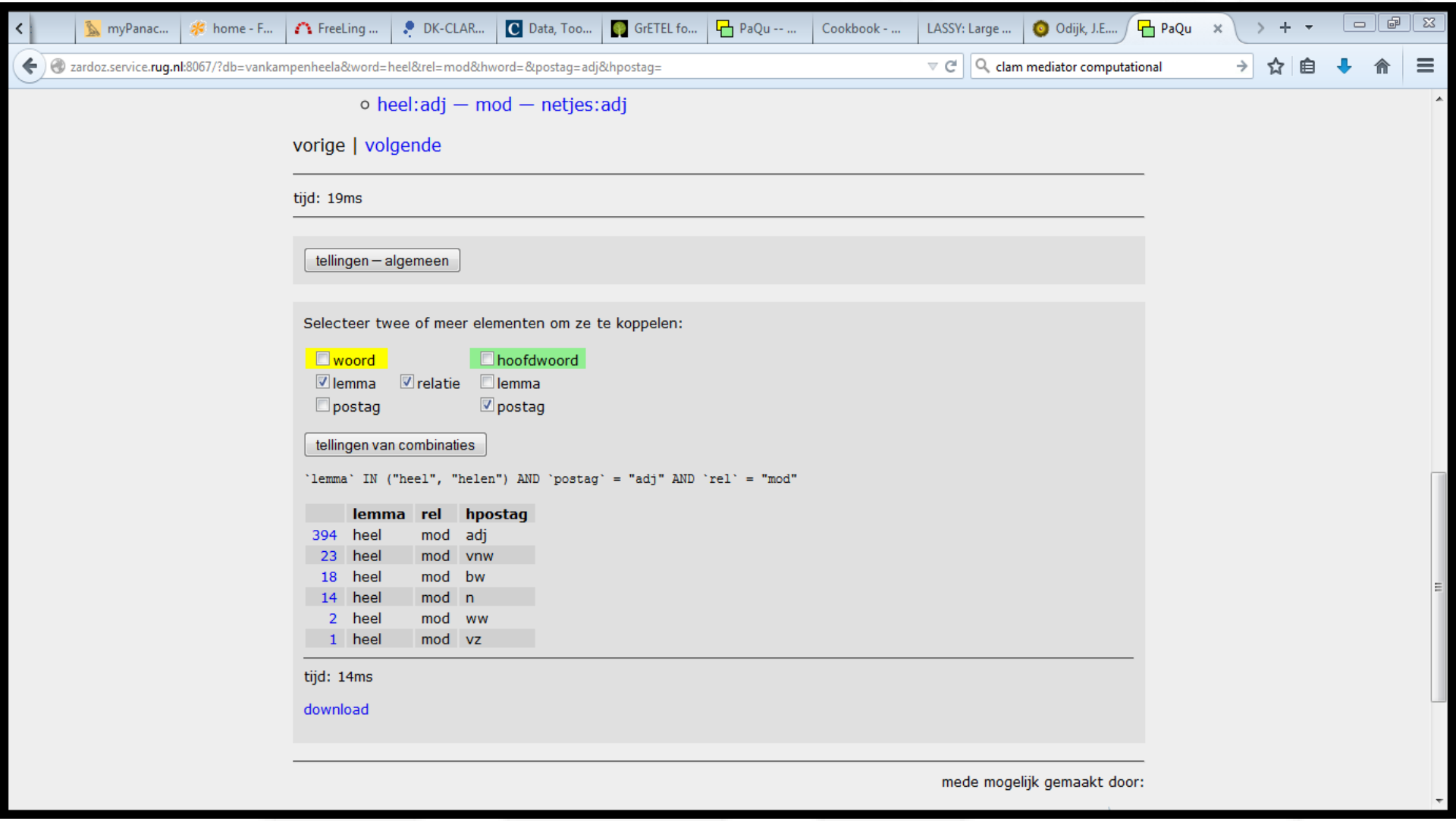
PaQu

- Small Experiment:
  - I selected the child-directed sentences containing *heel, erg, zeer* from the CHILDES Van Kampen corpus
  - Cleaned the text from annotations by a script
  - Loaded them to PaQu, and did an analysis.
  - Made a comparison with my manual annotation (see third lecture on the first day)- for the *mod* relation only.
  - Extended the manual annotation to cover multiple instances in one sentence (445 -> 452)
  - After comparison, I corrected the errors in my manual annotation (*revised manual annotation*)

PaQu


myPanac... home - F... FreeLing ... DK-CLAR... Data, Too... GrETEL fo... PaQu -- ... Cookbook - ... LASSY: Large ... Odijk, J.E.... PaQu
zardoz.service.rug.nl:8067/?db=vankampenheela&word=heel&rel=mod&hword=&postag=adj&hpostag=
clam mediator computational
Zoeken XPath Corpora Info
j.odijk@uu.nl Log uit
PaQu
corpus: VanKampenHeel – 445 zinnen
meer/minder
woord
heel
mod
adj
hoofdwoord
– postag –
help
Zoeken Wissen Reset
`lemma` IN ("heel", "helen") AND `postag` = "adj" AND `rel` = "mod"
1. het begin was heel goed .
heel:adj — mod — goed:adj
2. ik vind dat heel knap hoor , dat je dat durft te maken .
heel:adj — mod — knap:adj
3. ik wil iets heel leuks vinden .
heel:adj — mod — leuks:adj
4. dat is toch heel raar , daar staat helemaal niks op !
heel:adj — mod — raar:adj
5. die vind ik heel mooi , die rose , ja .
heel:adj — mod — mooi:adj

PaQu


zardoz.service.rug.nl:8067/?db=vankampenheela&word=heel&rel=mod&hword=&postag=adj&hpostag=
clam mediator computational
heel:adj — mod — netjes:adj
vorige | volgende
tijd: 19ms
tellingen – algemeen
Selecteer twee of meer elementen om ze te koppelen:
woord
lemma
relatie
postag
hoofdwoord
lemma
postag
tellingen van combinaties
`lemma` IN ("heel", "helen") AND `postag` = "adj" AND `rel` = "mod"
lemma rel hpostag
394 heel mod adj
23 heel mod vnw
18 heel mod bw
14 heel mod n
2 heel mod ww
1 heel mod vz
tijd: 14ms
download
mede mogelijk gemaakt door:

PaQu

- Comparison Alpino v. Manual v. Revised:

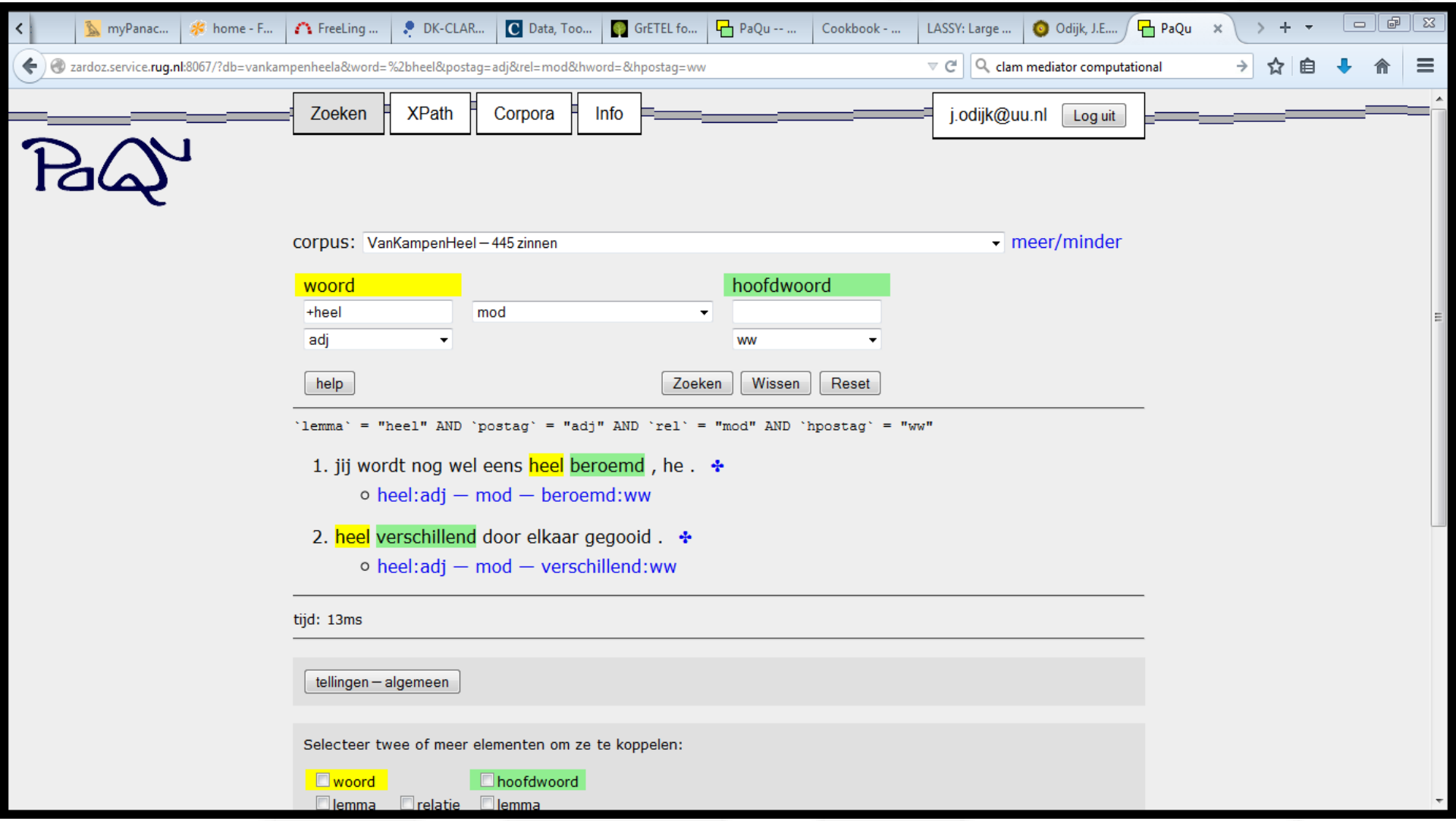
Case	Alpino	Manual	Revised
Mod adj	394	428	426
Mod vnw	23	--	--
Mod bw	18	--	--
Mod n	14	10	10
Mod ww	2	2	1
Mod vz	1	0	0
Predet n	0	1	2
predicative	0	7	8
unclear	0	4	5

PaQu

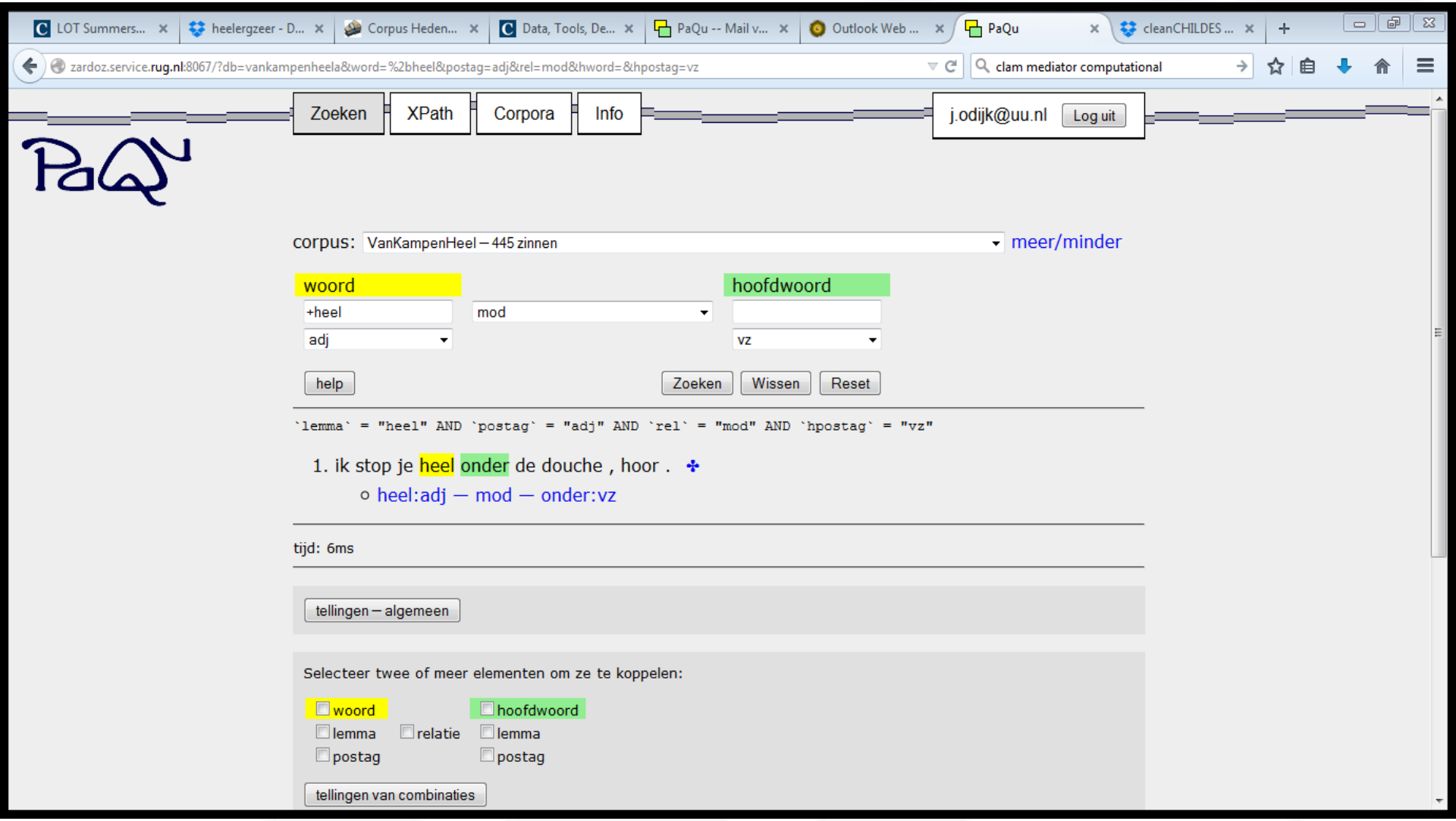
- Comparison Alpino v. Manual:

Count of examples	Column Labels						
Row Labels	**Amod**	**Nmod**	**NpreDet**	**pred**	**unclear**	**Vmod**	**Grand Total**
Amod	387	1		6	4		**398**
Bwmod	16					1	**17**
Nmod	6	6	1				**13**
Vmod	2						**2**
VNWmod	17	3		1			**21**
Vzmod						1	**1**
Grand Total	**428**	**10**	**1**	**7**	**4**	**2**	**452**

PaQu


myPanac... home - F... FreeLing ... DK-CLAR... Data, Too... GrETEL fo... PaQu -- ... Cookbook - ... LASSY: Large ... Odijk, J.E.... PaQu
zardoz.service.rug.nl:8067/?db=vankampenheela&word=%2bheel&postag=adj&rel=mod&hword=&hpostag=ww
clam mediator computational
Zoeken XPath Corpora Info
j.odijk@uu.nl Log uit
PaQu
corpus: VanKampenHeel – 445 zinnen
meer/minder
woord
hoofdwoord
+heel
mod
adj
ww
help
Zoeken Wissen Reset
`lemma` = "heel" AND `postag` = "adj" AND `rel` = "mod" AND `hpostag` = "ww"
1. jij wordt nog wel eens heel beroemd , he .
heel:adj — mod — beroemd:ww
2. heel verschillend door elkaar gegooid .
heel:adj — mod — verschillend:ww
tijd: 13ms
tellingen – algemeen
Selecteer twee of meer elementen om ze te koppelen:
woord
hoofdwoord
lemma
relatie
lemma

PaQu


LOT Summers...
heelergzeer - D...
Corpus Heden...
Data, Tools, De...
PaQu -- Mail v...
Outlook Web ...
PaQu
cleanCHILDES ...
zardoz.service.rug.nl:8067/?db=vankampenheela&word=%2bheel&postag=adj&rel=mod&hword=&hpostag=vz
clam mediator computational
Zoeken
XPath
Corpora
Info
j.odijk@uu.nl
Log uit
PaQu
corpus: VanKampenHeel – 445 zinnen
meer/minder
woord
hoofdwoord
+heel
mod
adj
vz
help
Zoeken
Wissen
Reset
`lemma` = "heel" AND `postag` = "adj" AND `rel` = "mod" AND `hpostag` = "vz"
1. ik stop je heel onder de douche , hoor .
heel:adj — mod — onder:vz
tijd: 6ms
tellingen – algemeen
Selecteer twee of meer elementen om ze te koppelen:
woord
lemma
postag
relatie
hoofdwoord
lemma
postag
tellingen van combinaties

PaQu

- Accuracy Alpino:
  - TolAlpino = VNWmod and BWmod counted as Amod

Alpino v Manual	0,87
Alpino v Revised	0,88
TolAlpino v manual	0,94
TolAlpino v revised	0,96

PaQu

- Initial Experiment is promising
  - Automatic parsing gives reasonable results for a small experiment involving CHILDES data
    - But it concerns adult speech
    - It mostly concerns a very local grammatical relation
- Many extensions are still needed
  - CHILDES CHAT format (and many other formats) should be natively supported, including cleaning
  - Search and analysis of data and all metadata should be supported

Other systems

- I am not aware of any, except perhaps:
  - TüNDRA - the *Tübingen aNnotated Data Retrieval Application*

Overview

- Enrich your own resources (TTNWW)
- Enrich your own resources (other applications)
- Search in your own enriched resources (AutoSearch)
- Search in your own enriched resources (PaQu)
- **Conclusions**

Conclusions

- TTNWW enables linguists to enrich their own data (for Dutch)
- There is an increasing number of such applications for other languages, in large part thanks to CLARIN
- Applications to upload such data into search engines are under development
- Initial results of a small experiment are promising

Thanks for your Attention!